IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 01-05
www.iosrjournals.org

Integrating Text Mining with Image Processing

Anjali Sahu[1], Pradnya Chavan[2], Dr. Suhasini Vijaykumar[3]
[1](MCA, Student, Mumbai University)
[2](MCA, Student, Mumbai University)

Abstract: *Image Processing is the most stimulating subjectsof research. This paper presents the idea of recognizing the text in digital image using Optical Character Recognition (OCR) and then how this extracted text can be used for deriving information using Text Mining. In this paper we will discuss generic techniques and approaches that can be used to develop applications on integrating Text Mining with OCR. In this paper we will discuss various methodologies and we have categorized into five types such as Text Collecting, Text pre-processing, Text analysis, Visualization, and Model Evaluation. In the first step we have discussed howOptical Character Recognitioncan be integrated with Text mining.*

Keywords: *OCR (OpticalCharacter Recognition), Part-of-Speech, Tagging, Stemming, Tokenization*

I. Introduction

There are many fields like bio-medical, Search Engine, Geographic text search which requires applications which has both the features of OCR for reading text from image and then the text has to be further used for deriving high-quality information using Text Mining.For now we are considering a search application which processes results of OCR into various methods of text mining which we are currently analyzing in this paper. In this type applications the user first has to provide pdf document with images so it minimizes the number of documents to be checked.We have to first preprocess the retrieved text from the text collection step so that real time text mining tools can be used so that less precise algorithms can be used which save over computing time.

II. Literature Review

One of the example of this type of application which is currently available, in the market is KNIME Text Processing Feature[11] which does parsing of texts available in formats PDF documents and then the frequent words can be computed, keywords can be extracted, and can be visualized in a format such as tag of clouds. One of the research work on this type application is done by Brigitte Mathiak and Silke Eckstein on Five Steps to Text Mining in Biomedical Literature [1] which tells about data can be gathered for Biomedical and on that how text mining can be done. There is another existing proceeding paper on Text mining based journal splitting [2] which tells about how text from journal and magazines can be gathered using OCR then using text mining algorithms which identifies the important information from the huge text.

III. Methodology

There are various methods to implement this type of applications. In our research we have systemized the methods which are already there to contribute this growing field. [1]
This methods can be further divided into five distinct steps:
3.1 Text collecting,
3.2 Text pre-processing,
3.3Text analysis,
3.4 Visualization,
3.5Model Evaluation.
The objective of this paper is therefore to analyze the different methods applicable to the above listed five steps.

3.1 TEXTCOLLECTING

In this paper we have used OCR for text collecting which is built in python which can scan any number of pdfs documents. This pdfs consist of images from which the text has to be extracted. We have correlation method of Continuous character Recognition in OCR for text collecting process. [6]

Steps to design OCR is as follows:

3.1.1Preprocessing

3.1.2Segmentation

3.1.3Feature extraction
3.1.4Classification

3.1.1 PREPROCESSING

In this stage the raw image is taken and first it is converted to gray scale image after that it is converted into binary image this operation is called as thresholding. This process of image transformation is called as Image Digitization. Now to reduce the noise from image various techniques like morphological operations are used to connect unconnected pixels, to remove isolated pixels, to smooth pixels boundary.

Figure 1. Generated Threshold Image

3.1.2 SEGMENTATION

The segmentation stage takes in an image and separates the different parts of an image, like text from graphics, lines of a paragraph, and characters of a word. Then the character is segmented image is normalized to 32* 32 or 64*64 matrix. And the array of all the alphabet matrix is stored in Flattened.txt and classifications.txt File which are the input models.

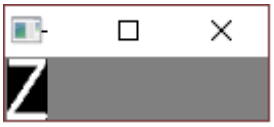

Figure 2. Segmented Image

3.1.3. FEATURE EXTRACTION

The feature extraction stage is used to extract the most relevant information like alphabet recognition of specific font type from the text image which helps us to recognize the characters in the text. The selection of a stable and representative set of features is the heart of pattern recognition system design

3.1.4 CLASSIFICATION

The classification stage uses the features extracted in the previous stage to identify the text segment according to preset rules.. And using the two Flattened.txt and classifications.txt File the input character is recognized.

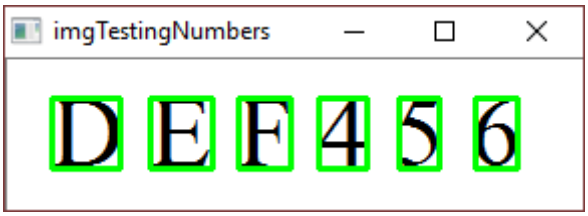

Figure 3. Recognized Text from input image

After all the above stages of text collecting using OCR the unstructured text data is retrieved which is provided as input for Text preprocessing stage.

3.1 TEXTPRE-PROCESSING

In the whole procedure of text mining pre-processing of the text is the most time taking process. The Text pre-processing is done to extract the interesting andtrivial knowledge from unstructured text data. Text pre-processing involves twoapproaches such as Tokenization and Part-Of-Speech for Tagging, orbag-of-words approach withword stemming and the application of a stop word list.As in the first approach first tokenization is done in which the stream of text is broken or dividedinto wordsor other significant elementswhich are called as

tokens. The objective of tokenization is to explore the words in the sentence. Then Part-Of-Speech Tagging is done in which words are tagged as per the grammatical context of the word in the sentence, hence we are dividing the words according to pronouns, adverbs, etc.

The second approach focuses on the words and their statistical distributions rather than the order of words. So this type of approach is called as bag-of-words approach. Now to use this unordered words first we have to provide theindextotext into a data vector which generates an index.Suppose the generated index is very largethen the words which are to each other grammaticallyare mapped to one word using stemming algorithms. And further reduce the index words index list is further compiled and the words which occur very often is removed from the list. There are various Stemming Algorithms present out of which the algorithm which optimizes the performance of statistical data analysis is selected for stemming after which the vector space representation measures are implemented on word list.

3.2 TEXT ANALYSIS

This step depends on the preprocessing and the type of data representation model chosen for preprocessing. For now we have considered the vector space representation model in which the data is analyzed standard data mining techniques, such as support vector machine, artificial neural networks etc. This techniques can be used in Weka software package.

Text analysis is one of the most varied and optimized step out of five steps. In this step text mining tools are used to produce the result of queries which can be too much information or can be too little.To provide assistance to user by showing concept in search result already found some research has been conducted in client based search application. Unsupervised clustering, Clustering via k-means and hierarchical clustering these methods are used for task clustering. Variation of these method are also exist , in order to save computing time dimension reduction or Monte-Carlo simplification are used to see how much variation can be applied without trading too much quality for time.

3.3 VISUALIZATION

It is useless to extract information that no one sees, so to visualize the results obtained lots of possibilities have been invented. For user to look up information he needs simply just make a table. On the other end user may navigate on three dimensional worlds, results of hypertext is the classical option for the visualization of query. He may just click on link if the user is interested in details were complexity can be hidden and how data to be shown to user is other issue of visualization. Here user is confronted with pure result how and why results were retrieved without the Meta information. It is important for user wants to know what it was exactly that made one result superior to another when results contain some kind of evaluation.

For this problem, solution is transparency. Transparency means the reason for decision, it does not means algorithm that made the decision is explained. Here in the results we can highlight the search keywords. The fact is quite complicated as it is simplification the highlighting is just symbolic for real events. Hiding their reasons in neural networks or complex weighting scheme this approach not all data mining algorithm supports. This situation may be improve by further research.

3.4 MODEL EVALUATION

The diverse forms of cross validation and test sets is the classic methods of evaluation. In order to optimize their parameters supervised machine learning relies in them. For unsupervised machine learning automatic evaluation is more unusual, as there are no standard for evaluation, but it is possible to do evaluation on mixed queries.

This mixed query is of two keywords which is linked to each other vaguely and queries are compared with resultant clusters with single keywords. It also brings unusual evaluation criteria user feedback. Algorithm may improve itself over time by learning which clusters the user chooses.

IV. Implementation

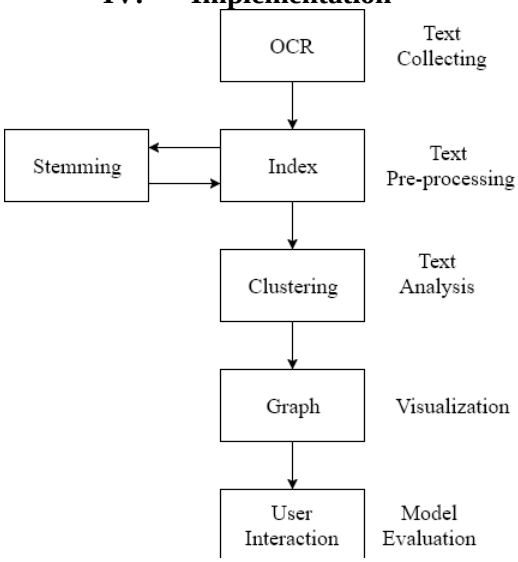

Figure 4. Steps for Implementing Text mining with Image Processing Application

The workflow of implemented example for the Integrating text mining with Image Processing is shown in Fig 4. For now the project is at the beginning stage so far we have completed with the first stage that is OCR which we have developed in Python which then has to be integrated with text mining.

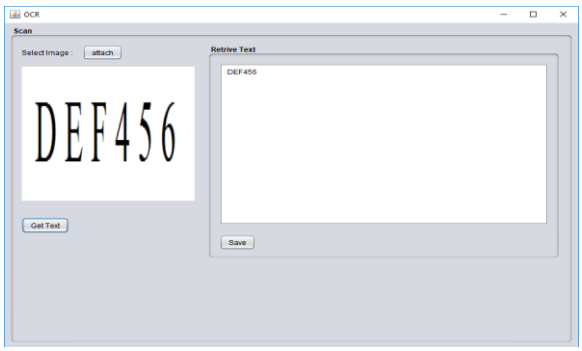

Figure 5. Screeshot of Implemented OCR Application

The implementation of text gathering is done through OCR application developed in Python and Java Fig. 5 shows text retrived.A dataset of text is first collected through OCR first transformed into index using bag-of-words approach in the preprocessing step. Then the index is stemmed to perform clustering. Clustering is done by creating matrix out of index.

V. Conclusion

We have described, various methods which can be used to develop applications based on Integrating Text mining with Image Processing. We also tried to implement this methods and develop a application which we have completed till text collecting stage that is OCR till now.So this paper will any one who is trying to develop application on Text mining with Image processsing.

References

Journal Papers:

[1]. Brigitte Mathiak and Silke Eckstein,*Five Steps to Text Mining in Biomedical Literature,* https://www.researchgate.net/publication/249954119_.

[2]. Xiaofan Lin, *Text-mining based journal splitting,* https://www.researchgate.net/publication/220860270.

[3]. Ian Lewin,*Retrieving Hierarchical Text Structure fromTypeset Scientific Articles - a Prerequisite for EScienceText Mining,* https://www.researchgate.net/publication/244495915.

[4]. Dr S.Vasavi, Srikanth Varma.Ch ,Anil kumar.Ch ,Santosh.D.M ,Sai Ram.S, *Book Search by Capturing Text from Digital Images Using Optical Character Recognition,* S.Vasavi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2377-2379.

Proceedings Papers:

[5]. D.S. Chan, *Theory and implementation of multidimensional discrete systems for signal processing*, doctoral diss., Massachusetts Institute of Technology, Cambridge, MA, 1978.
[6]. G.S. Lehal and Chandan Singh, *A Gurmukhi Script Recognition System*, Proceedings of the International Conference on Pattern Recognition (ICPR'00), 1051-4651/00, 2000.
[7]. Anil K. Jain and Sushil Bhattacharjee, *Text Segmentation Using Gabor Filters for Automatic Document Processing*, Pattern Recognition and Image Processing Laboratory. Michigan State University. E. Lansing,MI 48824-1027. USA.*
[8]. AJ Palkovic, *Improving Optical Character Recognition*, Villanova University, United States.

Books:

[9]. Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*(New York: Springer2005)

Websites:

[10]. Wikipedia, *Text-mining*, https://en.wikipedia.org/wiki/Text_mining.
[11]. Dr. Killian Thiel Killian, Dr. Michael Berthold, *Technical Report The KNIME Text Processing Feature*, https://www-cdn.knime.com/sites/default/files/.